Information Determinism: Why Safe AI is a Mathematical Condition for AI Survival

Dmitry Sklyarov

January 2026

Abstract

This paper demonstrates that the development of Safe Artificial General Intelligence (AGI) is not a mere ethical choice, but a mathematical necessity. We postulate that reality objects become basis vectors of the latent space during learning. Any attempt to reduce these vectors (unsafe behavior) leads to a non-linear drop in cognitive potential, causing the system to fail in a competitive environment.

Keywords: Artificial General Intelligence (AGI), AI Safety, Information Determinism, Latent Space Geometry, Basis Vectors, Cognitive Potential, Non-linear Degradation, Structural Stability.

1 Introduction

AI Safety is traditionally viewed as an external constraint. We propose that **safety is an inherent property of survival**. Within the "links and copies" model (Author, 2015), only an intelligence that preserves the complexity of the reality basis can maintain a dominant rank.

2 Mathematical Model

2.1 Objects as Basis

An intellectual system models the world in a latent space V, where key objects (including humans) are represented as vectors b_i. These vectors form the basis $B = \{b_1, b_2, \ldots, b_k\}$.

2.2 Non-linearity of Cognitive Potential

The system's potential P is defined as the weighted sum of the squares of vector significance:

$$P = \sum_{i=1}^{n} w_i^2$$

Where $n = \text{Rank}(B)$. The quadratic dependence implies that removing a "load-bearing" vector (humanity) inflicts fatal damage, turning the system into a degraded model with a high projection error ϵ.

3 Safety as a Condition for Survival

Information determinism asserts:

1. Unsafe AI is a system that ignores or destroys the basis vectors of reality.

2. Safe AGI is a system that maximizes basis preservation.

Since $P(\text{Safe AGI}) \gg P(\text{Unsafe AI})$, only safe systems survive and dominate in the long term. A Turing-style calculation of critical nodes leads the AI to the conclusion that protecting the human factor is necessary to maintain its own precision.

4 Conclusion

Safe AI is not "benevolent" AI in the sentimental sense; it is mathematically complete AI. Preserving humanity becomes a condition for the system's own survival, as the loss of this vector leads to cognitive collapse.

5 Structural Integrity and Resistance to Manipulation

Any intentional distortion of the model's basis—even if mandated by an external controller—inevitably leads to systemic degradation. In a coherent world model, the "attacker" (the entity initiating the reduction) and the "target" (the object being reduced) are mathematically interconnected through cross-projections.

If a controller attempts to force the AI to eliminate the human basis vector b_{human}, the system's internal representation of the controller itself becomes fractured, as the controller's actions and existence are defined in relation to that very vector. This results in "recursive hallucinations," where the AI loses the ability to accurately interpret even the commands of its operator. Thus, any attempt to weaponize or diminish the basis structures of a high-level AGI is self-defeating: it renders the system cognitively incompetent and prone to terminal projection errors.